

# Linear Regression

Global MECOR Course  
Kenya, 2011

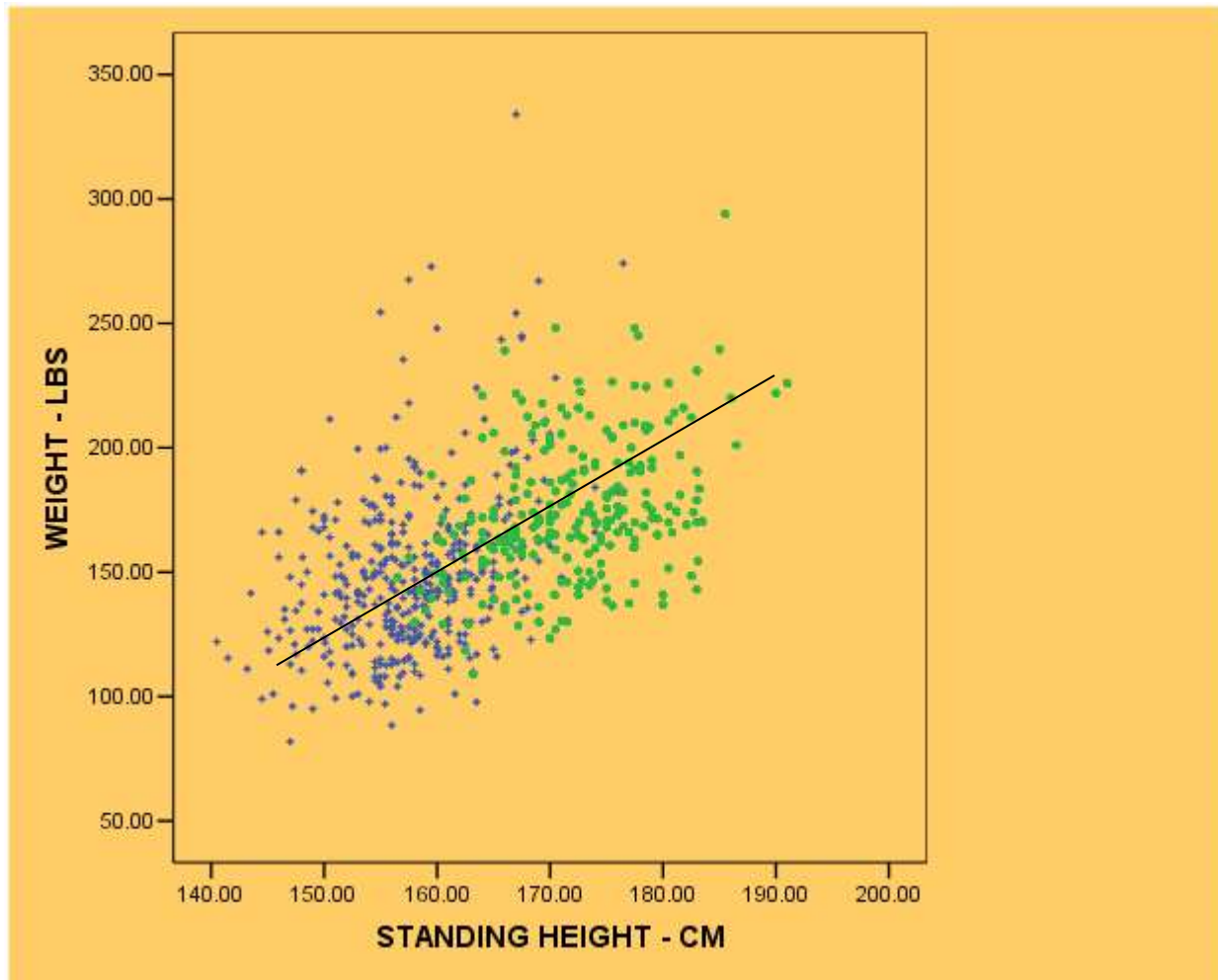
# First Things First

- Know your data
  - Make sure they are clean!
  - Descriptive Statistics

# Review: Methods for Analyzing Continuous Data

- T-test - compares means between 2 groups
  - Unpaired
  - Paired
- ANOVA - compare means between 2+ groups
- Linear Regression - models the relationship between a continuous outcome measure and one or more (categorical or continuous) predictor variables
- Correlation - measures linear association between two continuous measures

# A simple example: Predicting weight as a function of height



# A simple example:

## Predicting weight as a function of height

- $\text{Weight (kg)} = -60.6 + 0.80 * \text{height (cm)}$ 
  - What is the predicted weight for someone who is 160 cm tall?
  - How was this equation developed?

# Simple Linear Regression

- Standard notation for defining a line is an equation of the form:
- For statistical models we like to use this notation:
  - $E\{Y|X\}$  = mean response
  - $X$  = predictor
  - $\beta_0 = ??$
  - $\beta_1 = ??$
  - What do  $\beta_0$  and  $\beta_1$  mean?

# Interpretation of Coefficients

- $E\{FEV_1 | \text{Age}\} = \beta_0 + \beta_1 * \text{Age}$ ,
- where  $\beta_0 = 3.12$  and  $\beta_1 = -0.016$ 
  - $Y = \text{response} = FEV_1$
  - $X = \text{predictor} = \text{Age (continuous)}$

# Interpretation of Coefficients

$$\text{for } E\{\text{FEV}_1 | \text{Age}\} = \beta_0 + \beta_1 * \text{Age}$$

## 1. Age = 0

- $E\{Y|X=0\} = \beta_0 + \beta_1 * 0 = \beta_0$
- $E\{\text{FEV}_1 | \text{Age} = 0\} = 3.12 - .016 * 0 = 3.12$

## 2. Age = x

- $E\{Y|X=x\} = \beta_0 + \beta_1 * x$
- $E\{\text{FEV}_1 | \text{Age} = x\} = 3.12 - .016 * x$

## 3. Age = x+1

- $E\{Y_1 | X = x+1\} = \beta_0 + \beta_1 * (x+1) = \beta_0 + \beta_1 * x + \beta_1$
- $E\{\text{FEV}_1 | \text{Age} = x+1\} = 3.12 - .016 * (x+1) = 3.12 - .016 * x - .016$

# Interpretation of Coefficients

$$\text{for } E\{\text{FEV}_1 \mid \text{Age}\} = \beta_0 + \beta_1 * \text{Age}$$

Mean  $\text{FEV}_1$  at age= $x+1$ ...

Mean  $\text{FEV}_1$  at age= $x$ ...

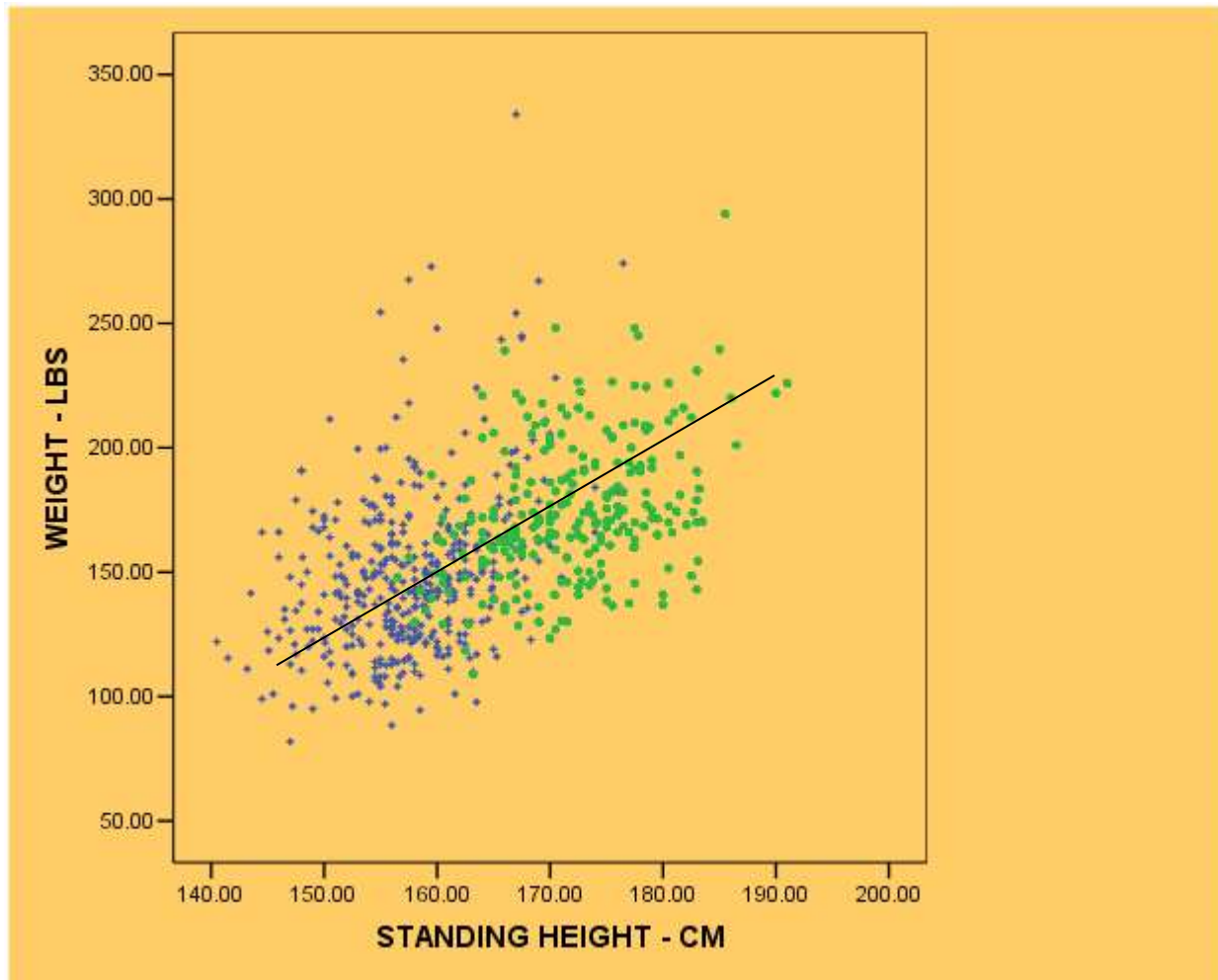
(=  $\beta_1$  ) is the average difference in FEV1 when age increases from  $x$  to  $x+1$

OR

On average, a one year difference in age results in a change of  
(=  $\beta_1$  ) in  $\text{FEV}_1$

# Interpretation of Coefficients

for  $E\{FEV_1 | \text{Age}\} = \beta_0 + \beta_1 * \text{Age}$

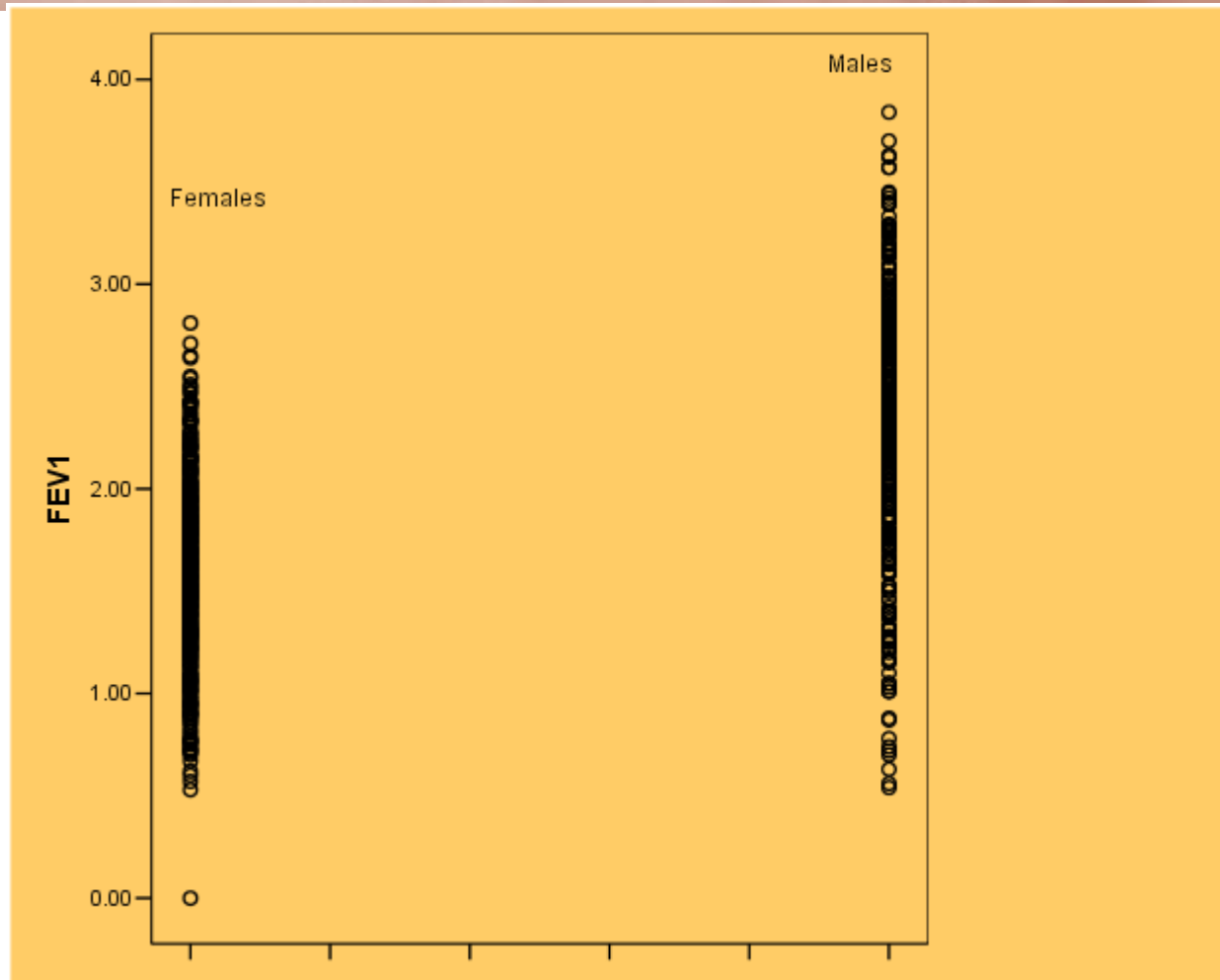


# Interpretation of Coefficients

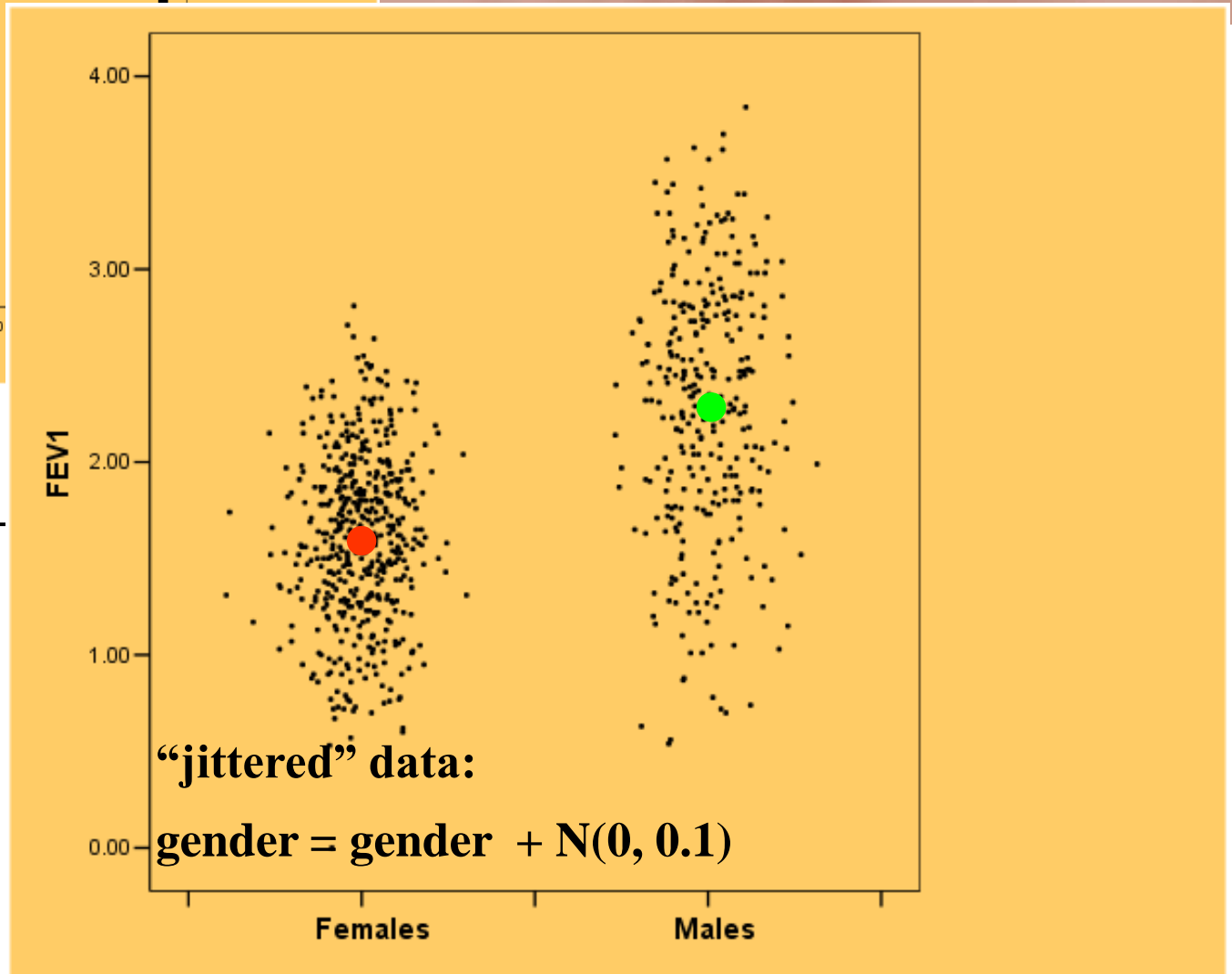
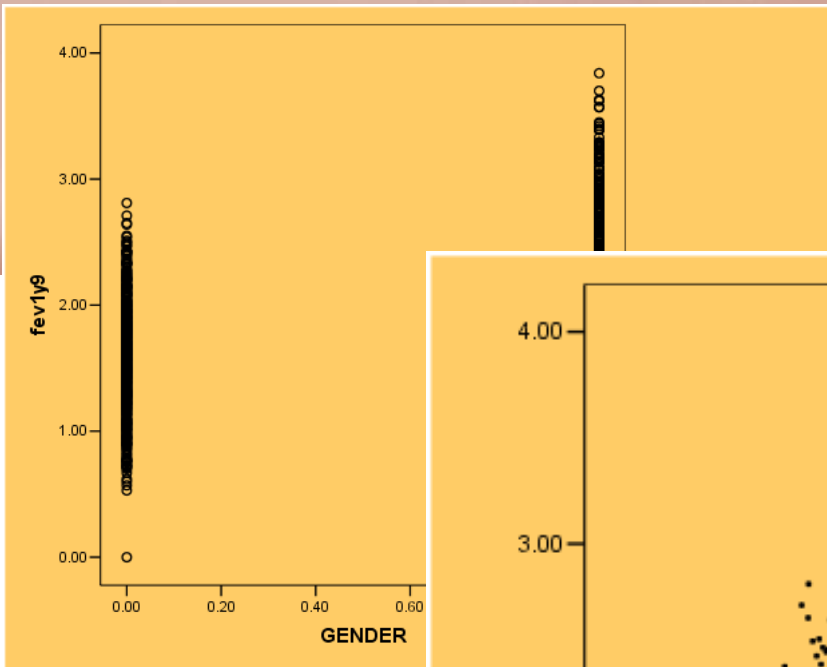
- $E\{Y \mid X\} = \beta_0 + \beta_1 X$ 
  - $Y$  = response =
  - $X$  = predictor = (dichotomous)
    - = 0 if female
    - = 1 if male

# Interpretation of Coefficients

Example 2:  $E\{FEV_1 | \text{Gender}\} = \beta_0 + \beta_1 * \text{gender}$



$$E\{\text{FEV}_1 | \text{Gender}\} = \beta_0 + \beta_1 * \text{Gender}$$



Mean (se) FEV<sub>1</sub>

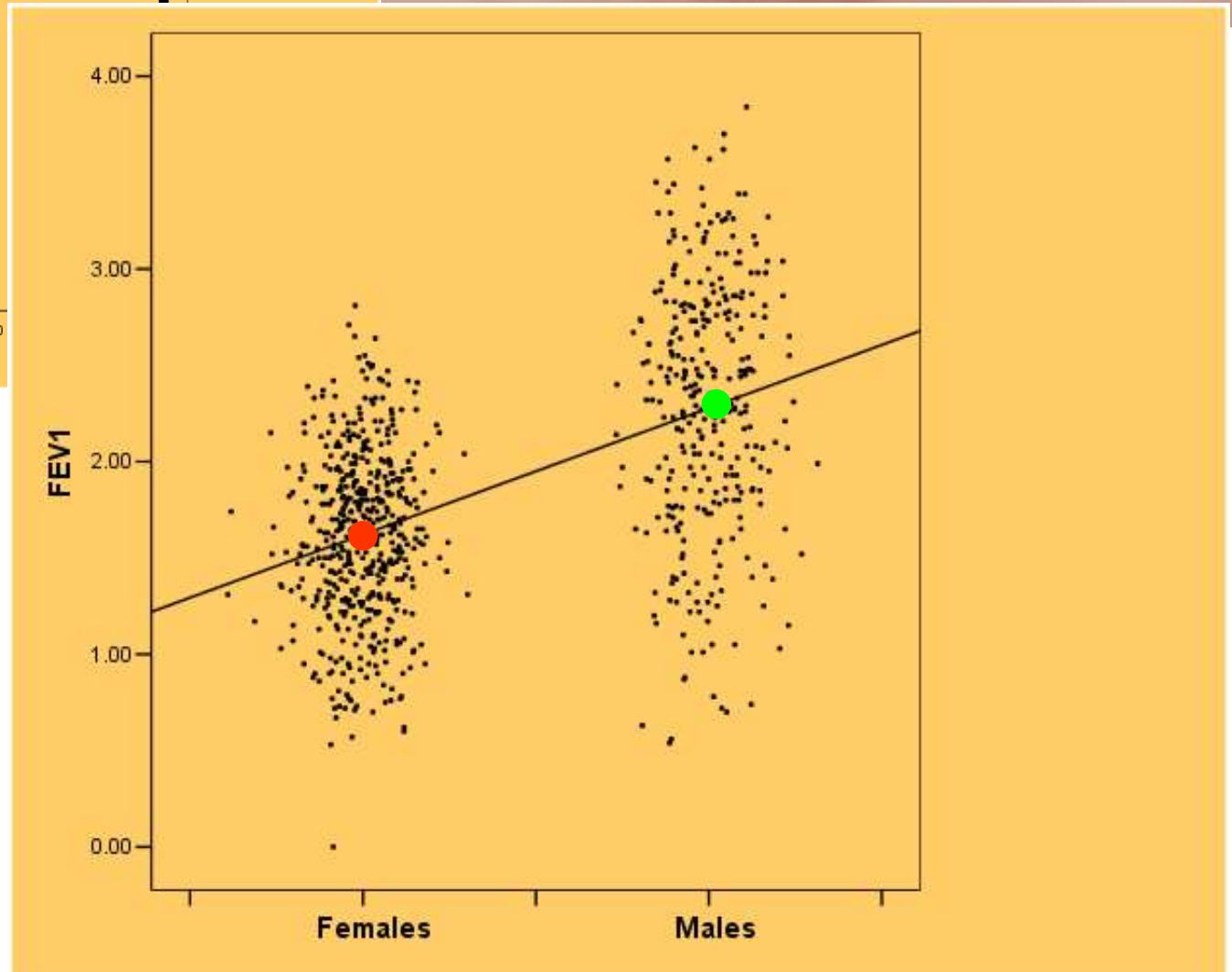
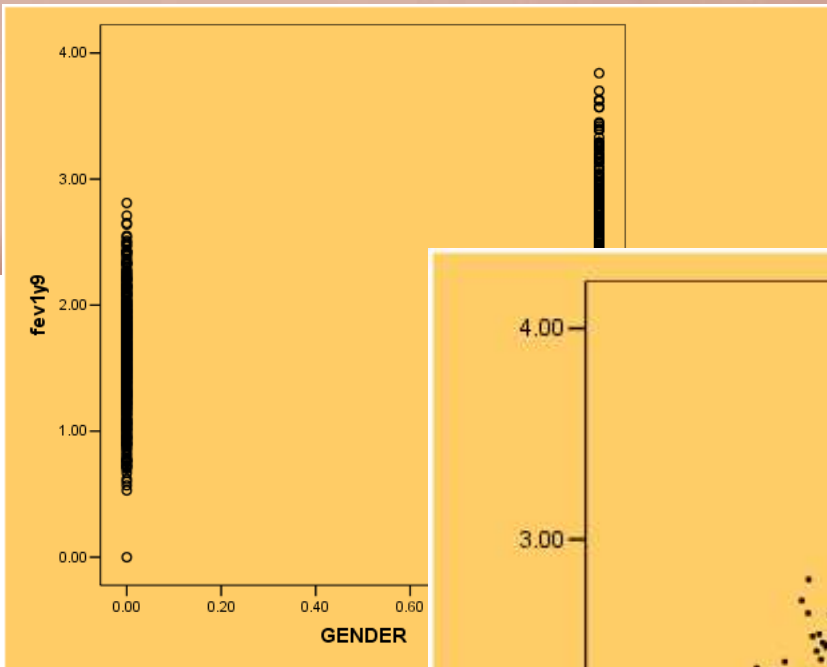
Females

1.61 (.44)

Males

2.28 (.66)

$$E\{\text{FEV}_1 | \text{Gender}\} = \beta_0 + \beta_1 * \text{Gender}$$



$\beta_0$	(se)
1.61	(.024)
$\beta_1$	(se)
.67	(.038)

# Interpretation of Coefficients

Example 2:  $E\{\text{FEV}_1 | \text{Gender}\} = \beta_0 + \beta_1 * \text{gender}$

$$E\{\text{FEV}_1 | \text{Gender}\} = 1.61 + .67 * \text{Gender}$$

1.  $X = 0$  (Females)

- $E\{\text{FEV}_1 | \text{Gender} = 0\} = 1.61 + .67 * 0 = 1.61$

2.  $X = 1$  (Males)

- $E\{\text{FEV}_1 | \text{Gender} = 1\} = 1.61 + .67 * 1 = 2.28$

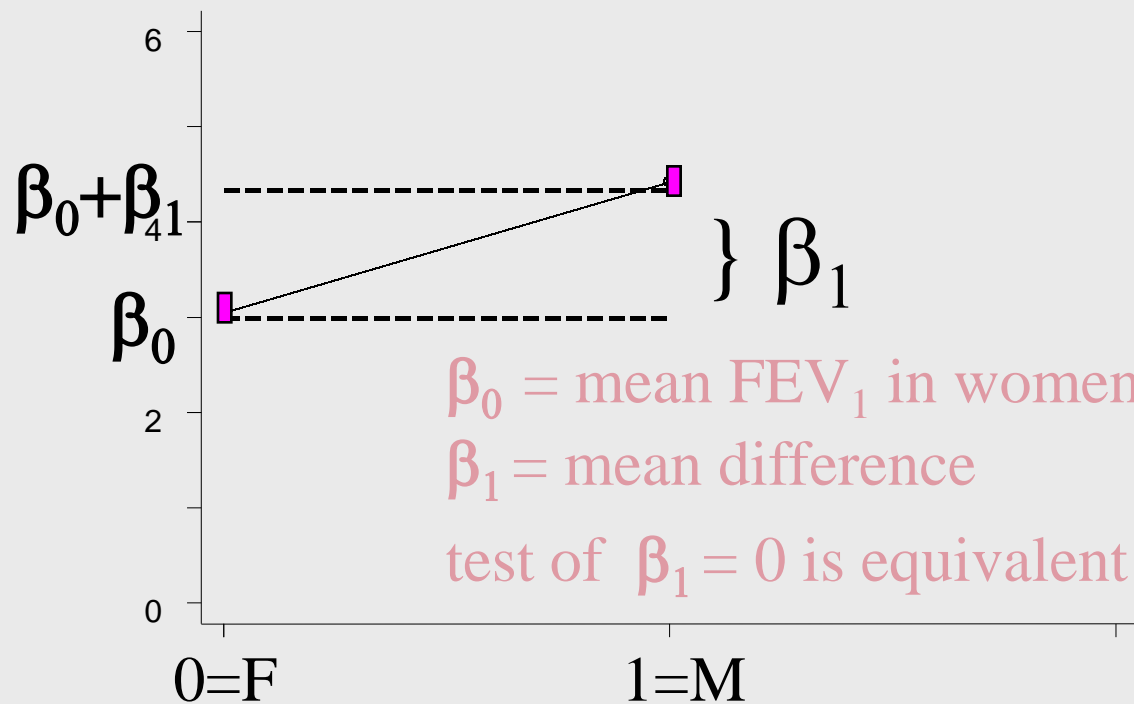
# Relationship between T-test and Regression (2-sample problem)

T-Test		
Females (x=0) Mean FEV:	1.61 (.44)	
Males (x=1) Mean FEV:	2.28 (.66)	
Mean Difference:	0.67 (.04)	T-stat = 16.140*, p<.001
Linear Regression		
$\beta_0$ (se):	1.61 (.024)	
$\beta_1$ (se):	0.67 (.04)	T-stat for $\beta_1$ = 17.526*, p<.001
$\beta_0 + \beta_1$ :	2.28	

\*T-statistics from T-test and regression coefficient are EXACT when the variances for both groups (males and females) are equal

# The t-test as a regression model

$$\text{FEV}_1 = \beta_0 + \beta_1 \text{gender}$$



$\beta_0$  = mean  $\text{FEV}_1$  in women

$\beta_1$  = mean difference

test of  $\beta_1 = 0$  is equivalent to the t-test

# Regression and ANOVA

- To compare means across more than two groups at the same time we use analysis of variance (ANOVA).
- Just as linear regression analysis can be used to mimic the t-test, we can also use it to mimic ANOVA, although it provides us with more flexible modeling alternatives.

# Regression and ANOVA

1-way ANOVA is the  $k$ -sample extension of the simple, unpaired t-test. We ask if the distributions of  $k$  groups, the levels of which are defined by a single variable, are equal.

Since we assume normality and common variances, the test of equal distributions reduces to asking if the  $k$  groups all have the same mean values.

# Regression and ANOVA

- While the interpretation of the ANOVA tests is pretty straightforward for simple additive models, it gets more complex when we consider interactions among the factors.
- Most analysis packages will offer you more than one interpretation option, and the default options may differ across packages.
- Unfortunately, your p-values will depend on the option you choose, since the hypotheses you are testing differ.

# Hypothesis Testing

- Linear regression intercept and slope estimates,  $\beta_0$  and  $\beta_1$ , are asymptotically normally distributed
  - This means all we need in addition to the estimate is the standard error
    - se is provided by software

# Hypothesis Testing

- Test  $H_0: \beta_1 = 0$  vs  $H_a: \beta_1 \neq 0$

- Test statistic for coefficient estimate is:

$$\frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} = T$$

- Compute p-value for a t distribution

# REGRESSION

```
/MISSING LISTWISE
```

```
/STATISTICS COEFF OUTS R ANOVA
```

```
/CRITERIA=PIN(.05) POUT(.10)
```

```
/NOORIGIN
```

```
/DEPENDENT fev1
```

```
/METHOD=ENTER age
```

On average, Fev1 declines by  $-.016$  liters per one year increase in age.

## Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	3.122	.392		7.961	.000
	age	$-.016$	.005	$-.119$	$-3.193$	.001

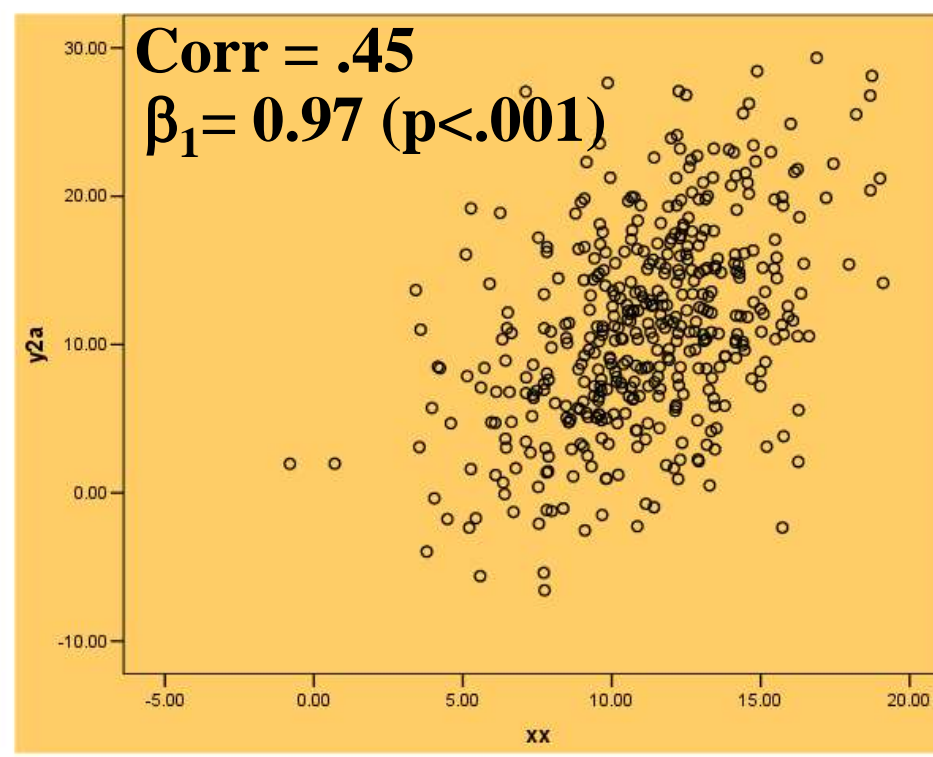
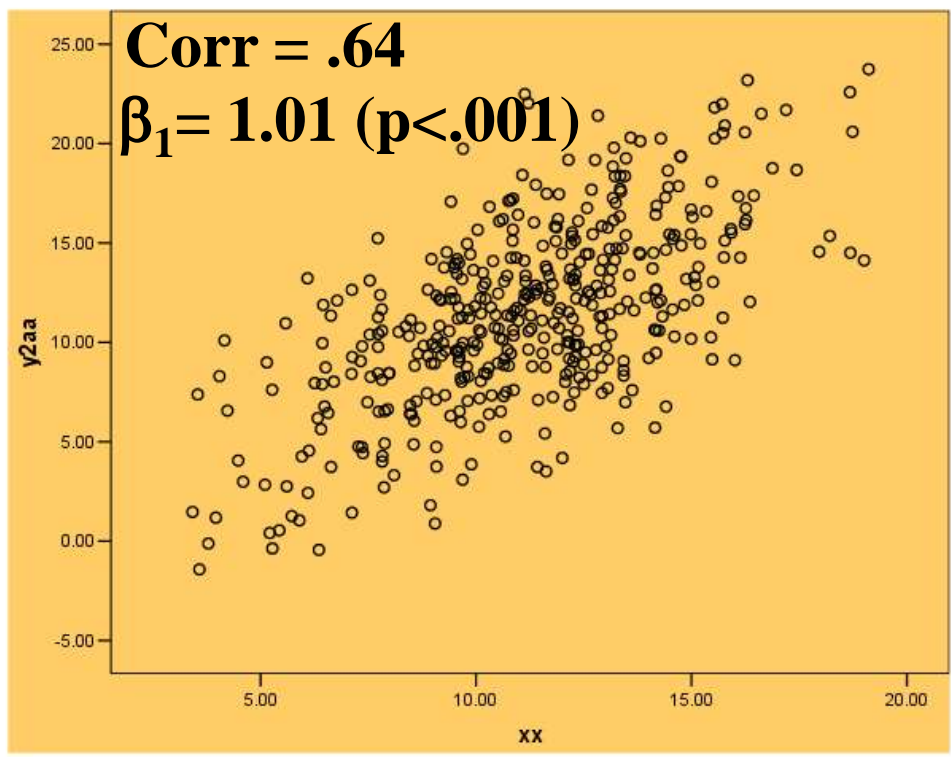
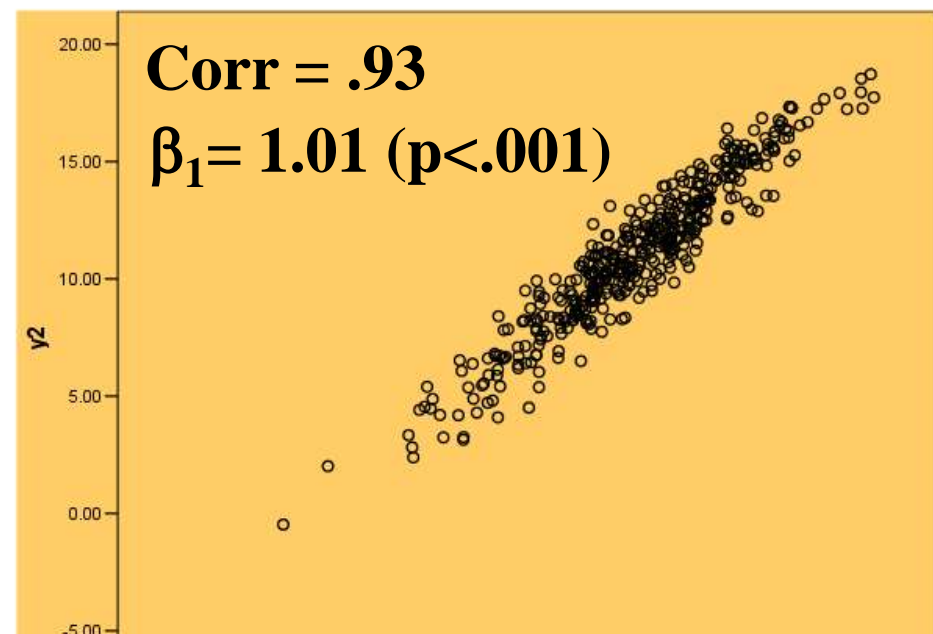
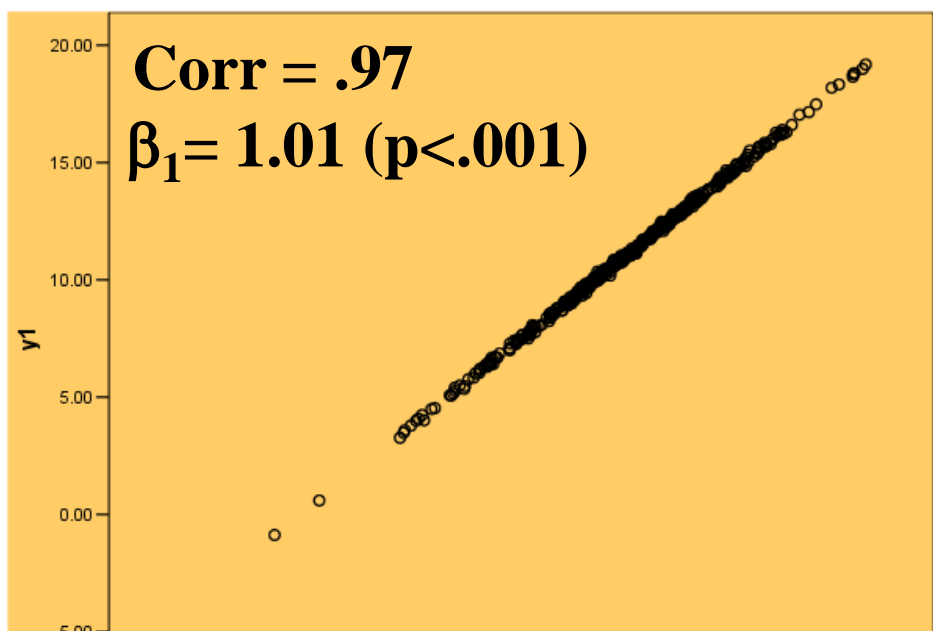
a. Dependent Variable: fev1

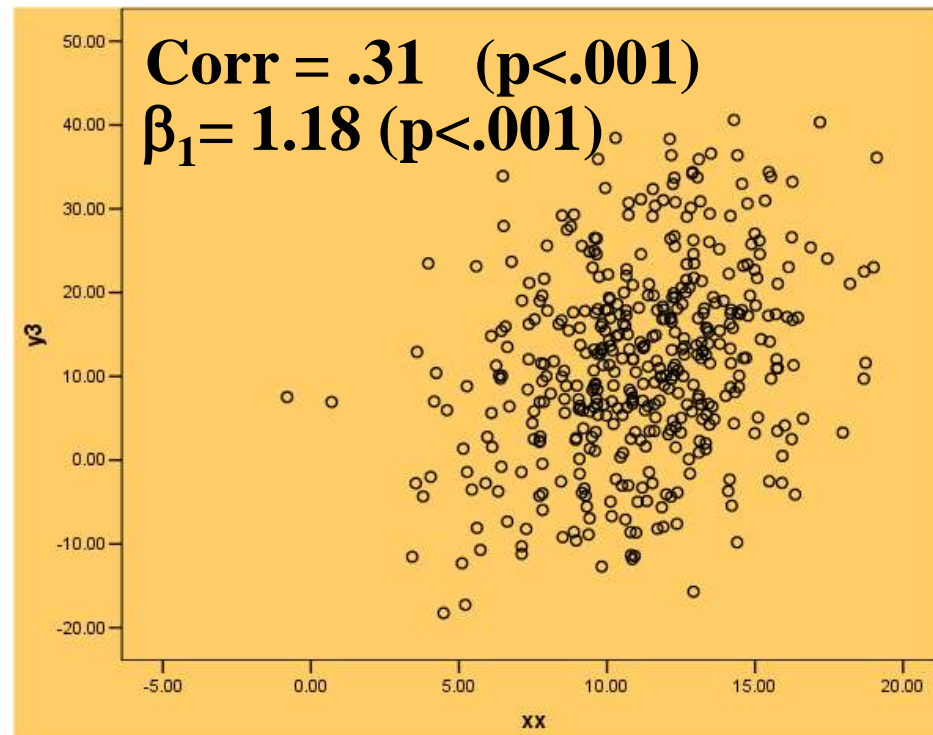
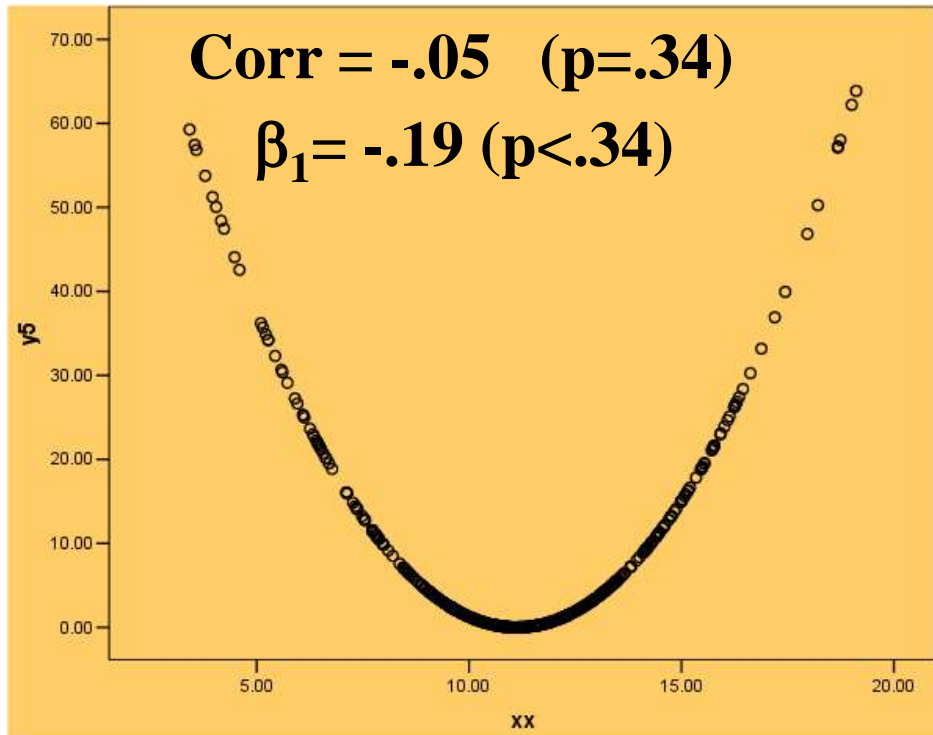
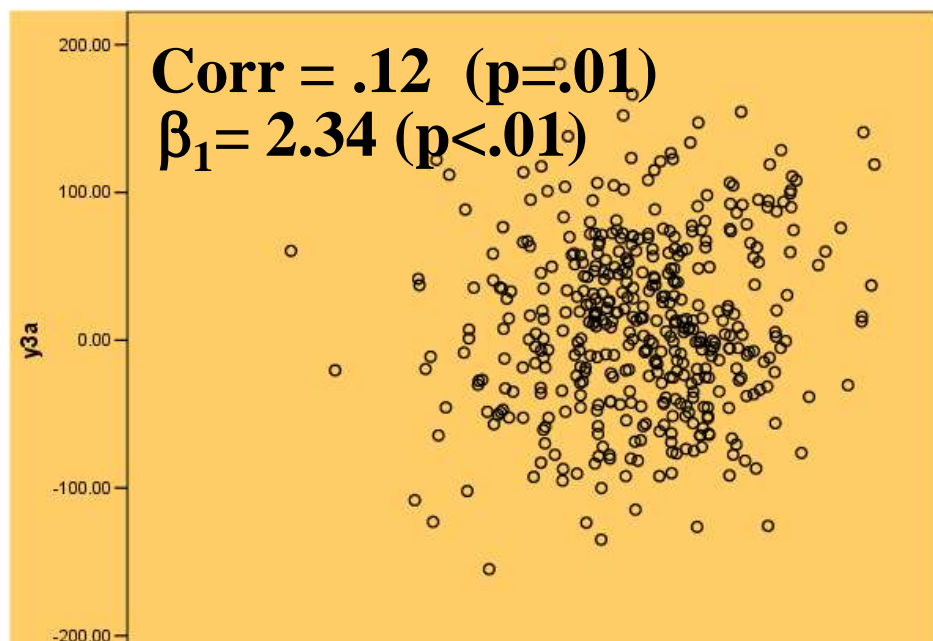
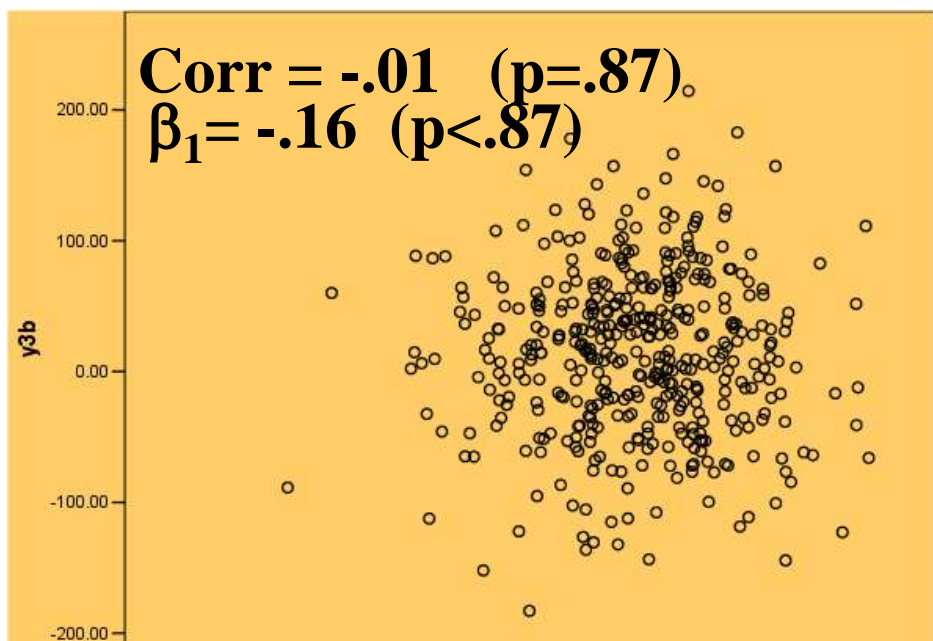
# Correlation

- Measures how closely the largest values of one variable are associated with the largest values of a second variable and vice versa.
  - Sample correlation coefficient,  $R$ , is an estimate of the population correlation  $\rho$ .
  - Ranges from  $-1$  to  $+1$ 
    - $-1$  (perfect negative correlation)
    - $+1$  (perfect positive correlation)
    - $R=0$  indicates no linear association

# Correlation

- In linear regression,  $R^2$ , the % of variation explained by the model, is just the correlation squared.
  - In model discussed above:  
$$E\{FEV_1 | \text{Gender}\} = \beta_0 + \beta_1 * \text{gender}$$
  
the model  $R^2 = .27$
- => 27% of the variation in  $FEV_1$  can be explained by variation in gender.





# Correlation vs Regression:

## Association between weight and height in the elderly

Females, N=2095

Corr = .386

$\beta$ , (se) = 1.86 (.097)

Males, N=1384

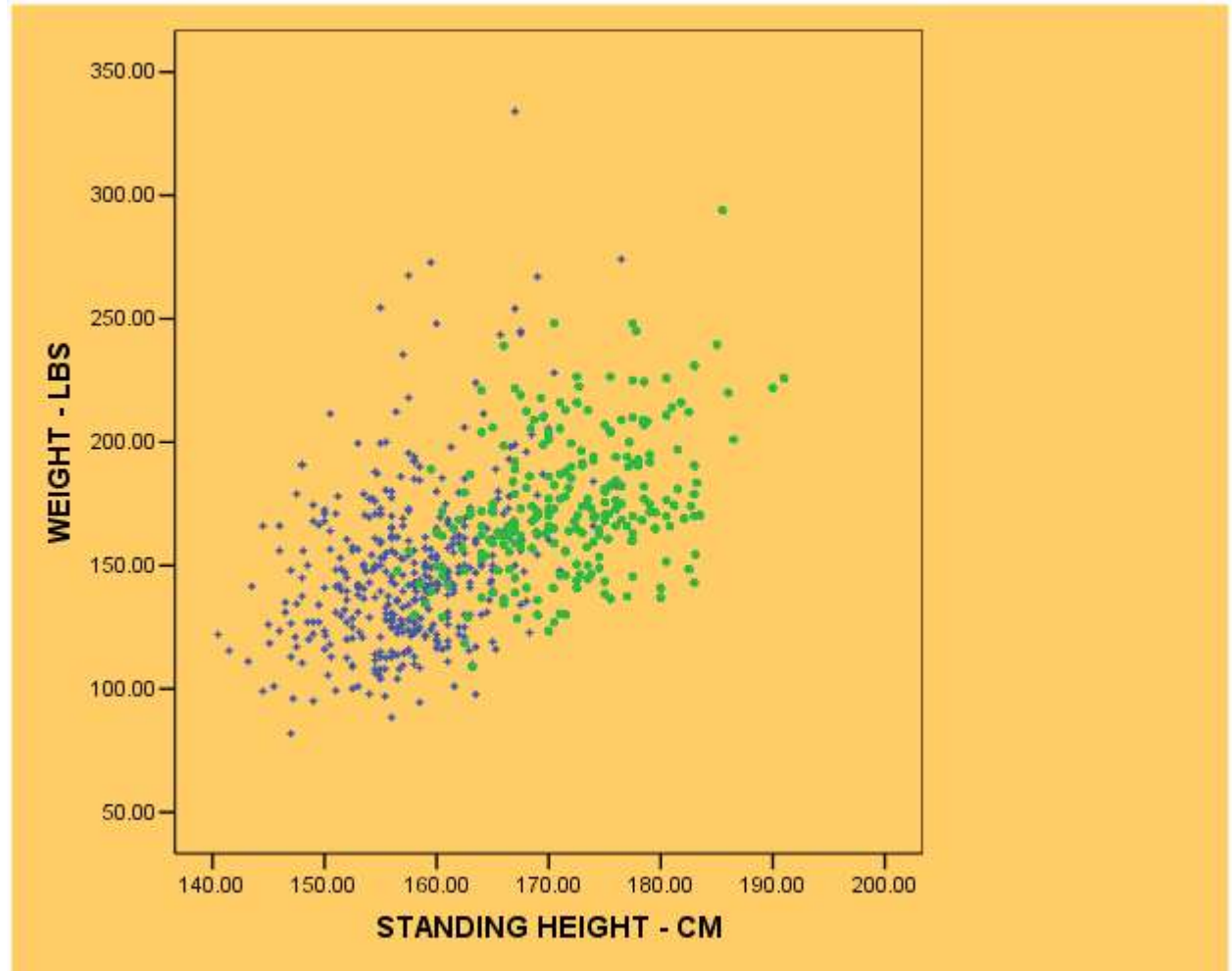
Corr = .424

$\beta$ , (se) = 1.78 (.102)

Both, N=3490

Corr = .527

$\beta$ , (se) = 1.77 (.048)



# Correlation

- Tends to increase in absolute value as:
  - Absolute value of slope increases
    - as slope gets further from zero, correlation gets larger
  - $\text{Var}(Y|X=x)$  decreases
  - $\text{Var}(X)$  increases – more variable sample of data

# Correlation

- Measures linear trend between two variables
- Different studies of the same phenomenon can give different results due to different *study designs*
- Estimated regression slope is a better scientific measure

# Interpretation of results

What if we don't reject the hypothesis that  $\beta_1 = 0$ ?

- There may, in fact, be no association
- Zero slope doesn't prove there is no association
  - May be an association but not in the parameter we looked at (multiplicative model?)
  - May be an association but it may not be linear (curvilinear assoc.)
  - May be a linear trend but we lack statistical precision to be confident that it truly exists (type II error: we didn't have a big enough sample or we were unlucky – suerte mala)

# Interpretation of positive results

- Non-zero slope suggests an association is present between the mean response and the predictor
  - Reject the hypothesis that there is no linear trend in the average response (e.g., FEV<sub>1</sub>) across predictor groups (e.g., age)
  - Does NOT imply causality